CUNY SPS DATA 621 - CTG5 - HW4

Betsy Rosalen, Gabrielle Bartomeo, Jeremy O'Brien, Lidiia Tronina, Rose Koh

April 24th, 2019

Contents

Table 1: Data Dictionary

VARIABLE	DEFINITION	TYPE
TARGET_FLAG	car crash = 1, no car crash = 0	binary categorical response
TARGET_AMT	car crash cost = >0, no car crash = 0	continuous numerical response
AGE	driver's age - very young/old tend to be risky	continuous numerical predictor
BLUEBOOK	$ value of vehicle	continuous numerical predictor
CAR_AGE	age of vehicle	continuous numerical predictor
CAR_TYPE	type of car (6types)	categorical predictor
CAR_USE	usage of car (commercial/private)	binary categorical predictor
CLM_FREQ	number of claims past 5 years	discrete numerical predictor
EDUCATION	max education level (5types)	categorical predictor
HOMEKIDS	number of children at home	discrete numerical predictor
HOME_VAL	$ home value - home owners tend to drive more responsibly	continuous numerical predictor
INCOME	$ income - rich people tend to get into fewer crashes	continuous numerical predictor
JOB	job category (8types, 1missing) - white collar tend to be safer	categorical predictor
KIDSDRIV	number of driving children - teenagers more likely to crash	discrete numerical predictor
MSTATUS	maritial status - married people drive more safely	catogerical predictor
MVR_PTS	number of traffic tickets	continuous numerical predictor
OLDCLAIM	$ total claims in the past 5 years	continuous numerical predictor
PARENT1	single parent	binary categorical predictor
RED_CAR	a red car	binary categorical predictor
REVOKED	license revoked (past 7 years) - more risky driver	binary categorical predictor
SEX	gender - woman may have less crashes than man	binary categorical predictor
TIF	time in force - number of years being customer	continuous numerical predictor
TRAVTIME	distance to work	continuous numerical predictor
URBANCITY	urban/rural	binary categorical predictor
YOJ	years on job - the longer they stay more safe	continuous numerical predictor

1 DATA EXPLORATION

In the pursuit of determining relationships between car crashes, their costs, and factors that may play a role into each, a dataset containing 8,161 observations with 25 variables was explored, analyzed, and modeled. This data came from an auto insurance company with each observation representing one of their customers. Of the 25 variables, two were target variables (car crashes and car costs), and the other 23 were predictors. `TARGET_FLAG` is a binary variable where a value of 1 indicates that the customer has made a claim related to a car crash and a value of 0 indicates they have not. The other target variable, `TARGET_AMT`, is a continuous numerical variable whose value is the payout amount of a claim, if any. The remaining variables are split in their categorization; 13 are categorical and 10 are numerical.

This data was utilized to compose and evaluate several types of models with the following features:

- Logistic classification models that aim to predict the probability that a person crashes their car; and,
- Multiple linear regression models that aim to predict the amount of money it will cost if the person does crash their car.

The intended use case for these models is actuarial in nature: specifically, to calculate insurance rates commensurate with policyholders' (or policy applicants') potential risk levels based on attributes such as income, age, distance to work, tenure as customers, so on and so forth.

Inspection of the target variables reveals that where `TARGET_FLAG` has values of 0 (i.e., no claim), `TARGET_AMT` also has values of 0 (i.e., no payout), which is logically consistent.

Table 2: Summary statistics

	n	min	mean	median	max	sd
TARGET_AMT	8161	0	1504.3	0	107586	4704.0
AGE	8155	16	44.8	45	81	8.6
YOJ	7707	0	10.5	11	23	4.1
INCOME	7716	0	61898.1	54028	367030	47572.7
HOME_VAL	7697	0	154867.3	161160	885282	129123.8
TRAVTIME	8161	5	33.5	33	142	15.9
BLUEBOOK	8161	1500	15709.9	14440	69740	8419.7
TIF	8161	1	5.3	4	25	4.2
OLDCLAIM	8161	0	4037.1	0	57037	8777.1
MVR_PTS	8161	0	1.7	1	13	2.1
CAR_AGE	7651	0	8.3	8	28	5.7

Table 3: Summary statistics for Categorical Variables

	EDUCATION	JOB	CAR_TYPE	KIDSDRIV	HOMEKIDS	CLM_FREQ
	<High School:1203	Blue Collar :1825	Minivan :2145	0:7180	0:5289	0:5009
	Bachelors :2242	Clerical :1271	Panel Truck: 676	1: 636	1: 902	1: 997
	Masters :1658	Professional:1117	Pickup :1389	2: 279	2:1118	2:1171
	PhD : 728	Manager : 988	Sports Car : 907	3: 62	3: 674	3: 776
	High School :2330	Lawyer : 835	Van : 750	4: 4	4: 164	4: 190
	NA	Student : 712	SUV :2294	NA	5: 14	5: 18
	NA	(Other) :1413	NA	NA	NA	NA

1.1 Summary Statistics

Continuous and categorical variables were summarized separately for the sake of clarity.

`EDUCATION`, `JOB`, `CAR_TYPE`, `KIDSDRIV`, `HOMEKIDS`, and `CLM_FREQ` each comprise multiple categories. On the other hand, `PARENT1`, `SEX`, `MSTATUS`, `CAR_USE`, `RED_CAR`, `REVOKED`, `URBANICITY` are all binaries.

1.2 Variable Descriptions

1.2.0.1 KIDSDRIV

`KIDSDRIV` is a categorical predictor with values ranging from 0 to 4. It shows heavy skew, with most cars having no kid drivers (value of 0). Judging from the distribution, it appears that having a kid driver results in higher probability of making a claim.

1.2.0.2 AGE

`AGE` presents driver's age and shows a normal distribution centered around 45 years. Looking at the boxplot of age below, there does not appear to be a difference in the distribution between whether a claim is made or not. Accordingly, this `AGE` may not be helpful in determining the probability of making a claim.

1.2.0.3 HOMEKIDS

`HOMEKIDS` is a predictor describing number of children at home ranging from 0 to 5.

1.2.0.4 YOJ

`YOJ` is a predictor describing years on job. People who stay at a job for a longer time are believed to be safer drivers. Apart from those who are unemployed (values of 0), `YOJ` seems to show a normal distribution.

Table 4: Summary statistics for Binary Categorical Variables

	PARENT1	SEX	MSTATUS	CAR_USE	RED_CAR	REVOKED	URBANICITY
	No :7084	M:3786	Yes:4894	Commercial:3029	no :5783	No :7161	Urban:6492
	Yes:1077	F:4375	No :3267	Private :5132	yes:2378	Yes:1000	Rural:1669

1.2.0.5 INCOME

`INCOME` is a heavily skewed predictor variable, suggesting that outliers should be treated for modelling.

1.2.0.6 HOME_VAL

`HOME_VAL` is a home value predictor variable. In theory, home owners tend to drive more responsibly. The difference between owners and renters (values of 0) is visible in the summary statistics graph.

1.2.0.7 TRAVTIME

`TRAVTIME` is a predictor variable describing the distance to work. Long drives to work would suggest greater risk of an accident and claim. However, its graph shows a fairly normal distribution, such that this variable may not be helpful in determining the probability of making a claim.

1.2.0.8 BLUEBOOK

`BLUEBOOK` is a predictor variable describing the value of the car. The boxplot demonstrates that the lower value of the car, the higher chances of making a claim. It is conceivable that higher-priced cars are driven more carefully.

1.2.0.9 TIF

`TIF` describes how long the customer has been with the insurance company. Plots reveal that the longer the tenure of a policyholder, the lower the likelihood of a claim - i.e. safe drivers tend to remain so.

1.2.0.10 OLDCLAIM

`OLDCLAIM` is a predictor describing the value of claims made in the past 5 years. It is very heavily skewed as most policyholders do not make claims.

1.2.0.11 CLM_FREQ

`CLM_FREQ` is a predictor that describing the frequency of claims in the past 5 years. It suggests that those who have made a claim in the past 5 years are more likely to make another claim.

1.2.0.12 MVR_PTS

`MVR_PTS` is a predictor that describes motor vehicle record points. The rationale is that more traffic tickets suggests less safe driving and a higher likelihood of claims. It appears to be a highly significant variable as seen in boxplots.

1.2.0.13 CAR_AGE

`CAR_AGE` describes the age of the policyholder's vehicle. One value is -3, which must be an error - this is corrected to 0.

1.2.0.14 PARENT1

`PARENT1` indicates whether a policyholder is a single parent. This variable has been factorized and relabeled as `NumParents` to describe the number of parents.

1.2.0.15 SEX

`SEX` describes the gender of the driver. This variable has been factorized and relabeled as `MALE`, for which males receive a value of 1 and females a value of as 0. It does not appear to be significant variable in the boxplots below.

1.2.0.16 MSTATUS

`MSTATUS` describes the marital status of the policyholder. The rationale is that married people drive more safely. This variable has been factorized and relabeled as `Single`, for which married policyholders receive a value of 0 and unmarried a value o 1.

1.2.0.17 EDUCATION

`EDUCATION` describes the education level of the driver. This variable is factorized. It may be correlated with INCOME.

1.2.0.18 JOB

`JOB` describes the type of job the driver has. This variable is factorized. It may be correlated with INCOME. In theory policyholders with white collar jobs tend to drive more safely.

1.2.0.19 CAR_TYPE

`CAR_TYPE` describes type of car. This variable is factorized.

1.2.0.20 CAR_USE

`CAR_USE` describes how the vehicle is used. Commercial vehicles are driven more and may have an elevated probability of accidents and claims. This variable is factorized and relabeled as `Commercial`, for which a value of 0 means private use and a value of 1 means commercial use.

1.2.0.21 RED_CAR

`RED_CAR` indicates whether the color of the vehicle is red. Red vehicles, especially sports cars, are associated with riskier driving and likelihood of claims. This variable is factorized.

1.2.0.22 REVOKED

`REVOKED` describes whether a policyholders license has been revoked in the past 7 years. License revocation is associated with riskier driving. This variable is factorized. The boxplot reveals that policyholders who previously lost their license are more likely to file claims.

1.2.0.23 URBANICITY

`URBANICITY` describes whether driver lives in an urban area or a rural area. This variable has been factorized and relabeled as `URBAN`, for which a value of 0 means rural and a value of 1 means urban.

1.2.1 Summary Statistics Graphs

[GAB: This sentence needs rephrasing and a supporting chart... if the supporting chart is the next chart, then this needs to be/should be moved] Examining the dispersion of claims between variables, it looks like likelihoods are higher for drivers who are male, urban, blue collar, unmarried, or parents; as well as for those with commercial vehicles or a revoked license.

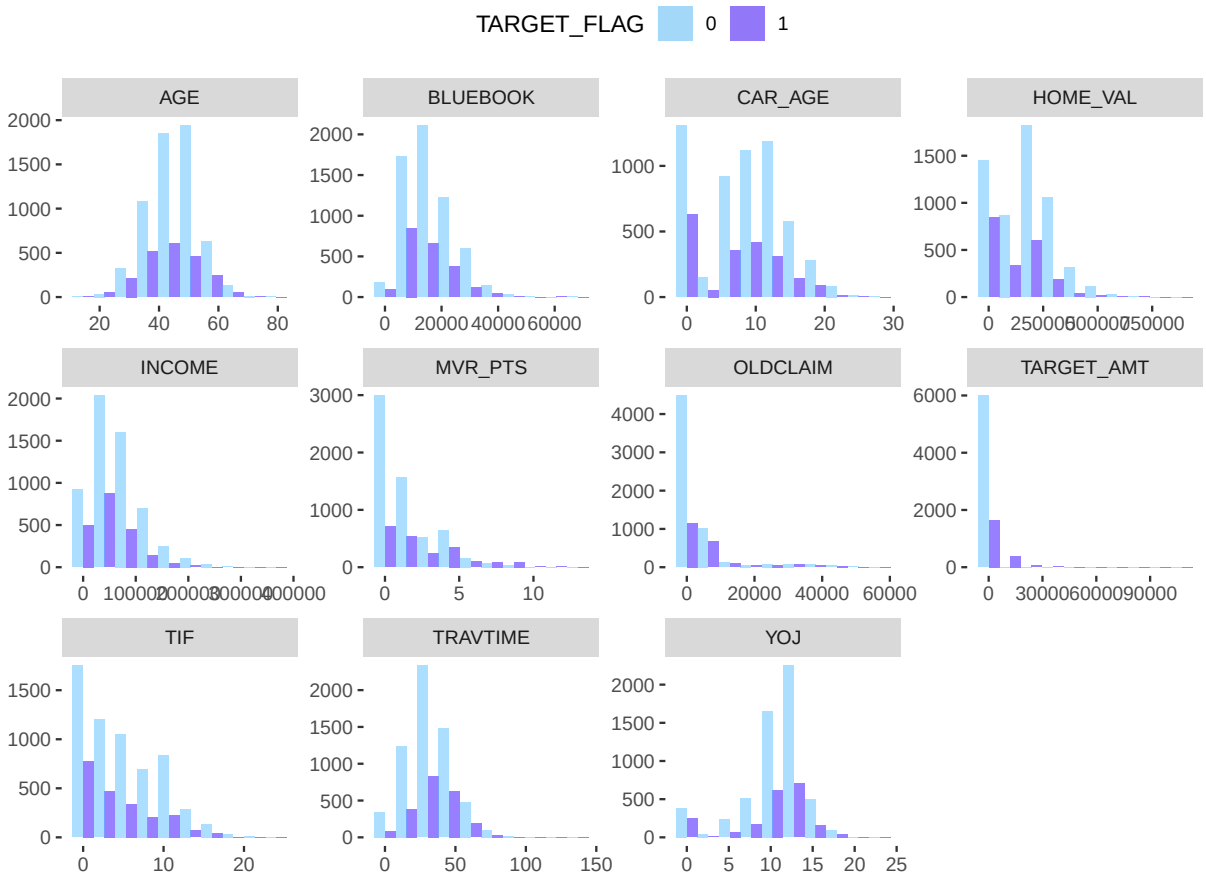


Figure 1: Numeric Data Distributions as a Function of TARGET_FLAG

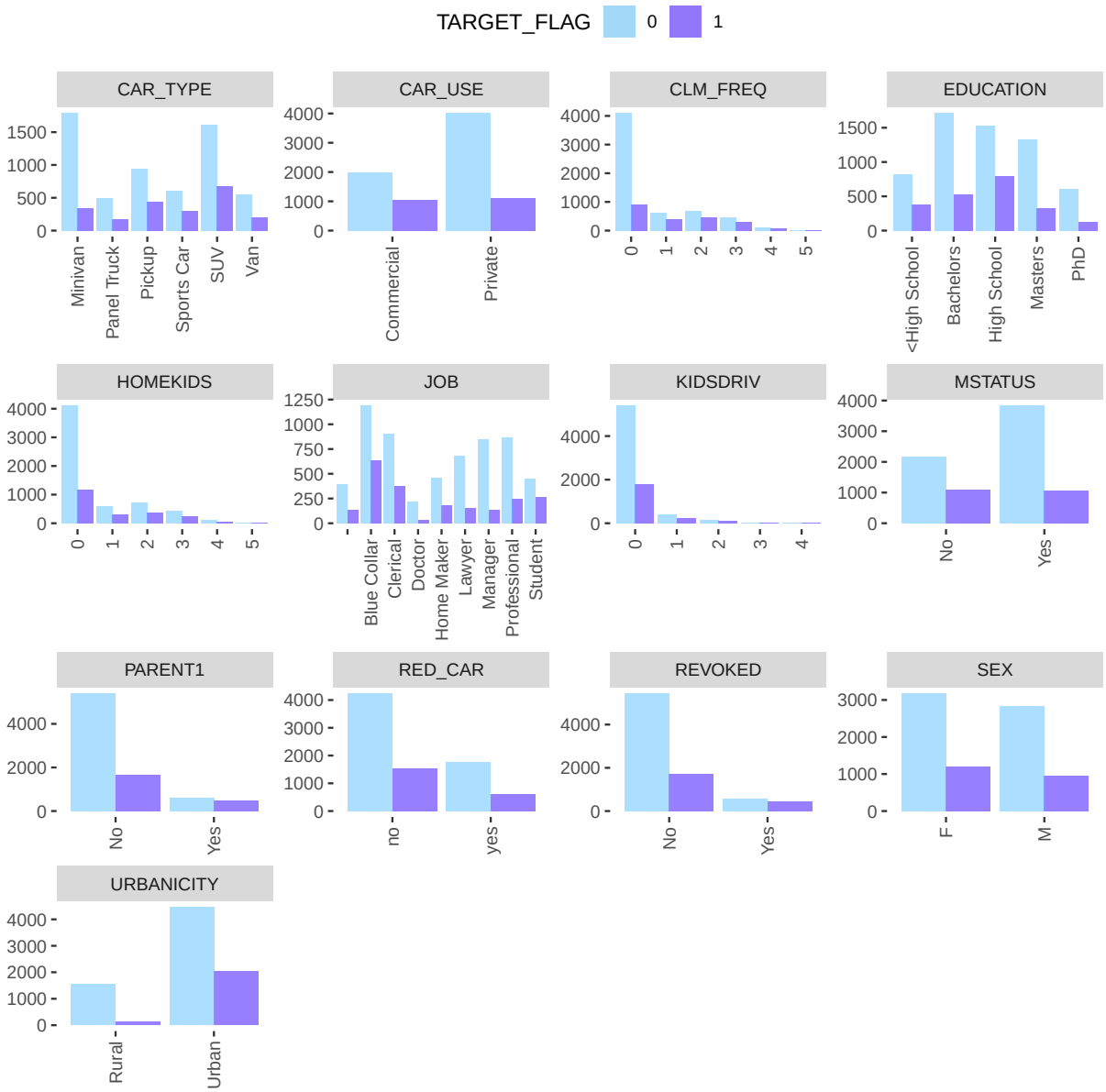


Figure 2: Categorical Data Distributions as a Function of TARGET_FLAG

The scale of the continuous variables' distributions are considerably different and difficult to visualize together.

Scaling the distribution based on the standard deviation reveals that outliers are very abundant for the continuous variables `OLDCLAIM`, `INCOME`, `TRAV_TIME`, `BLUEBOOK`, and to a lesser extent `HOME_VAL` and `TIF`. The variables that appear to have the most outliers are `OLDCLAIM`, `BLUEBOOK`, `TRAVTIME`, and `INCOME`. All of the variables show varying levels of skew save `YOJ` and `AGE` which appear the most normally distributed.

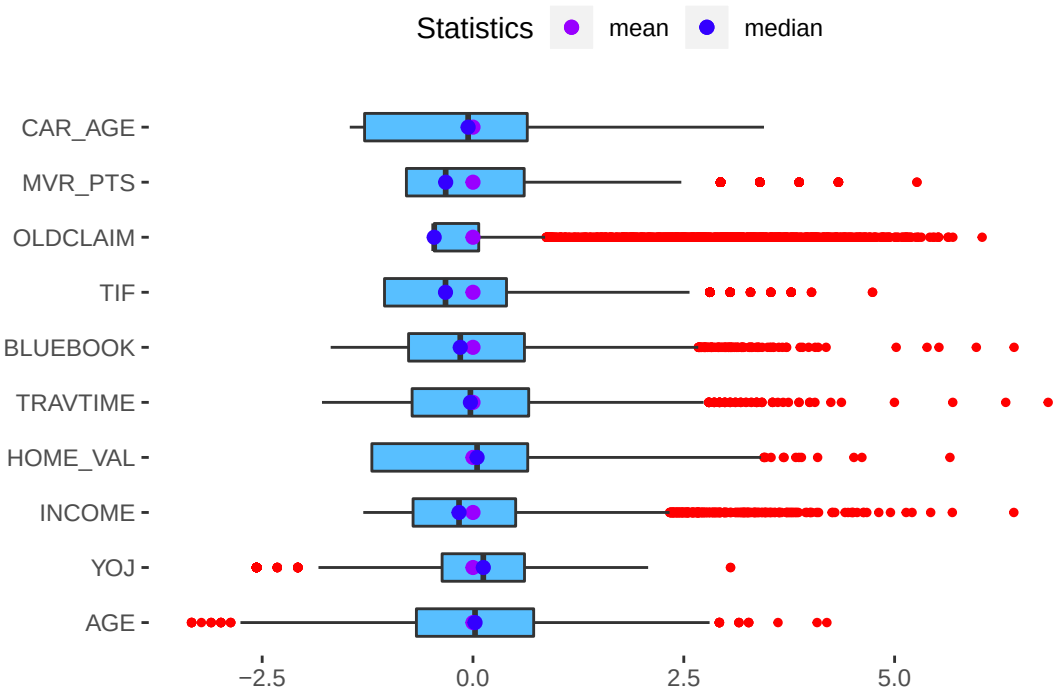


Figure 3: Scaled Boxplots

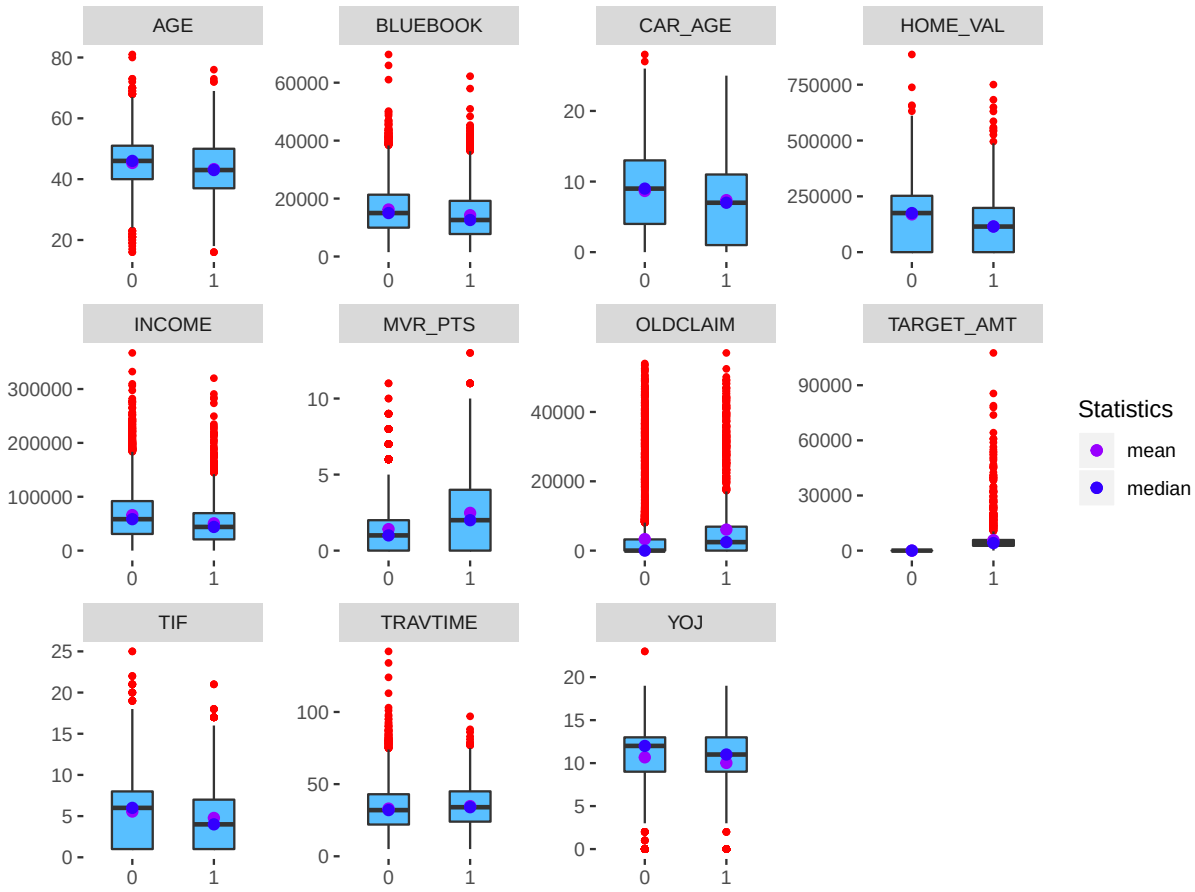


Figure 4: Linear relationship between each numeric predictor and the target

1.3 Linearity

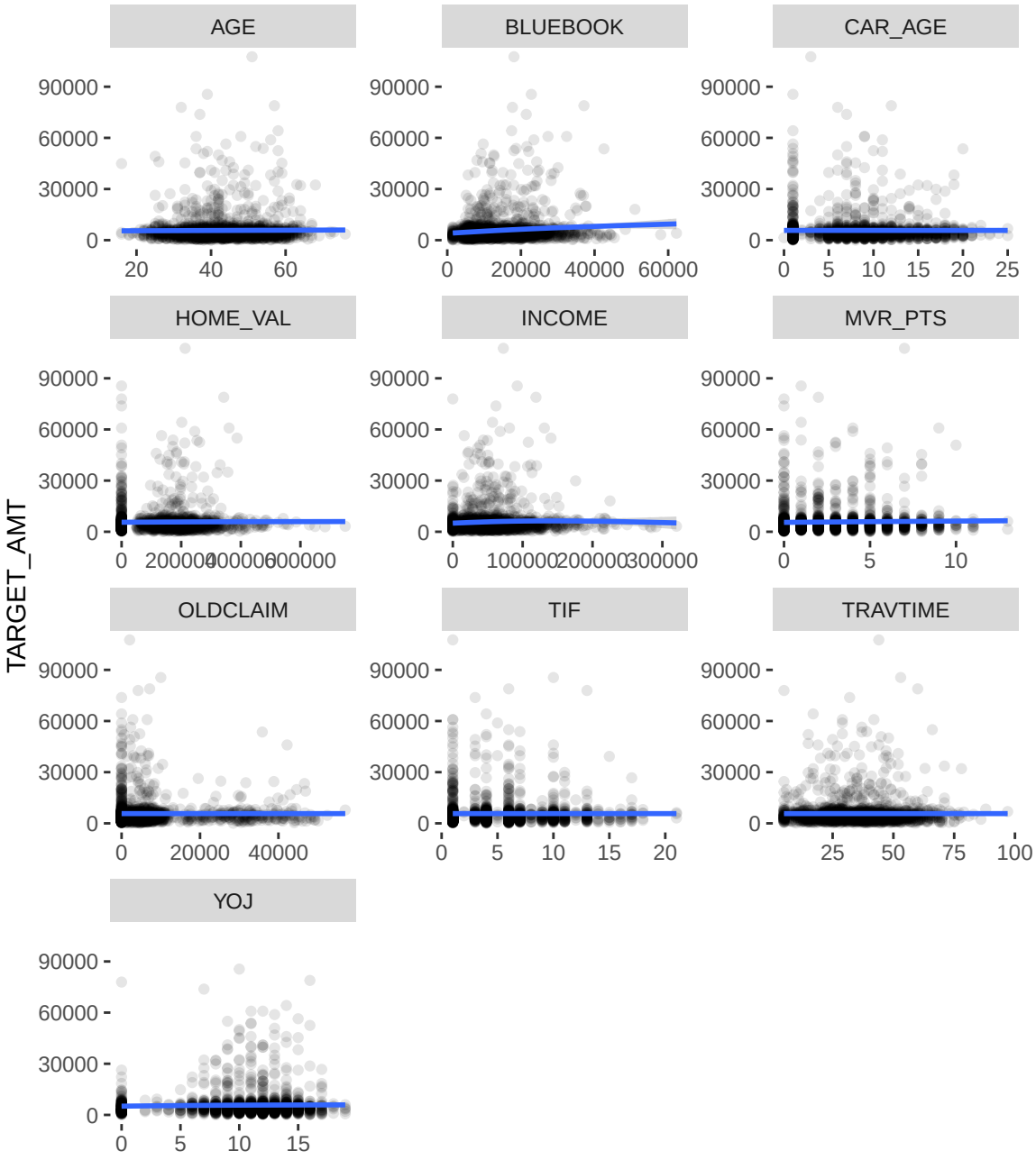


Figure 5: Scatter plot between numeric predictors and the TARGET_AMT, filtered for rows where TARGET_AMT is greater than 0

The plotted numeric predictors with their raw values fail to show any clear linear relationships with the `TARGET_AMT` except for the faintest of linearity in the `BLUEBOOK` variable.

[JO: DON'T THINK THE LOG TRANSFORM HAS HELPED - THINK WE SHOULD ADD]

[GAB: ADD WHAT? LOG DIDN'T HELP THOUGH YOU'RE RIGHT]

1.3.1 Log Transformed Data

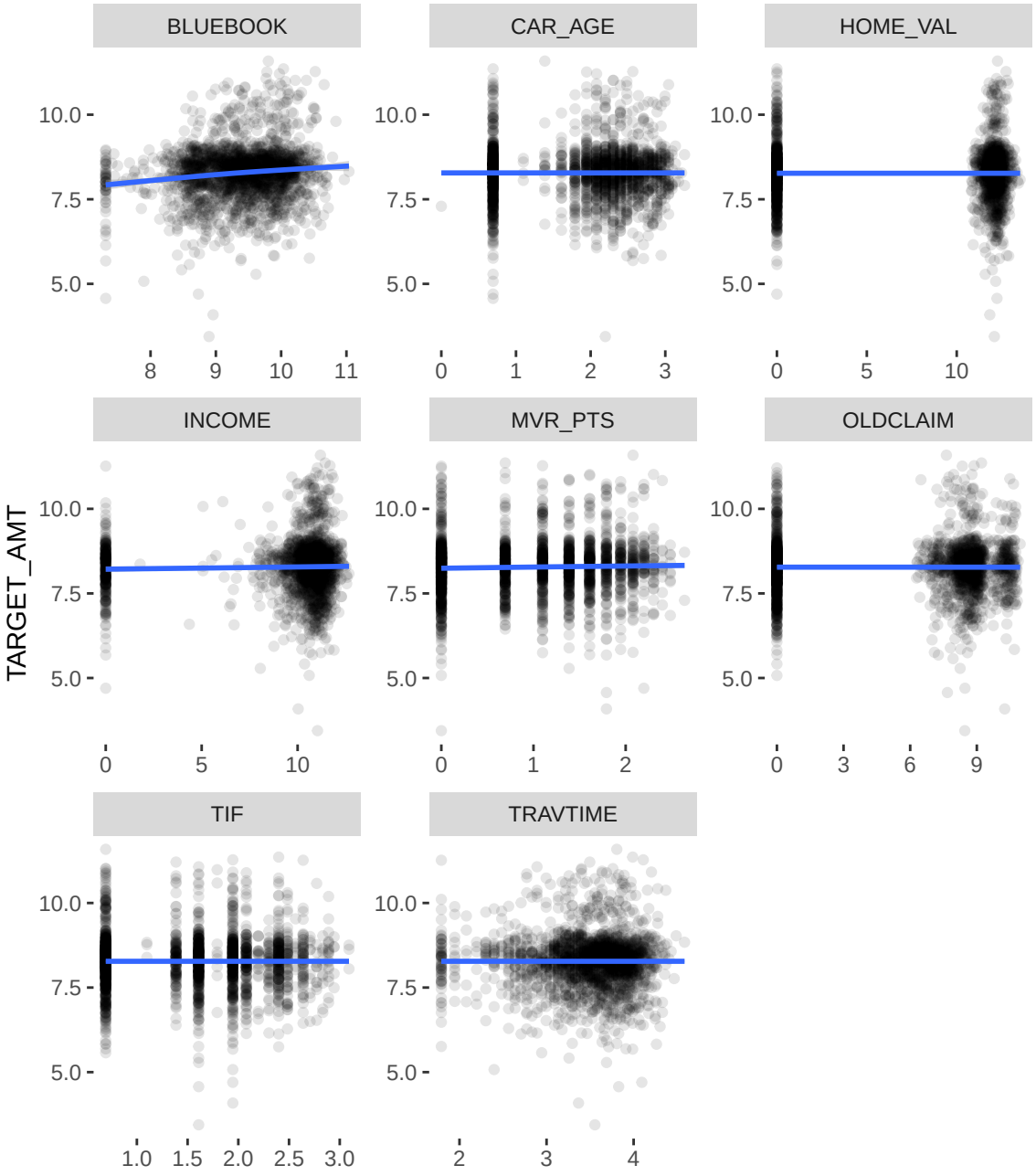


Figure 6: (#fig:f6.1)Scatter plot between log transformed numeric predictors and the log transformed TARGET_AMT filtered for rows where TARGET_AMT is greater than 0

In an attempt to distinguish the linearity of the variables alongside the `TARGET_AMT`, all numeric predictors and the `TARGET_AMT` underwent a log transformation. As a result, the linearity of `BLUEBOOK` became more apparent, but there was no obvious influence on the linearity of any of the other variables.

1.3.2 Box-Cox

Even though the linearity plots above don't show much improvement after a log transformation, a Box-Cox plot shows that a log transformation is recommended for the `TARGET_AMT`.

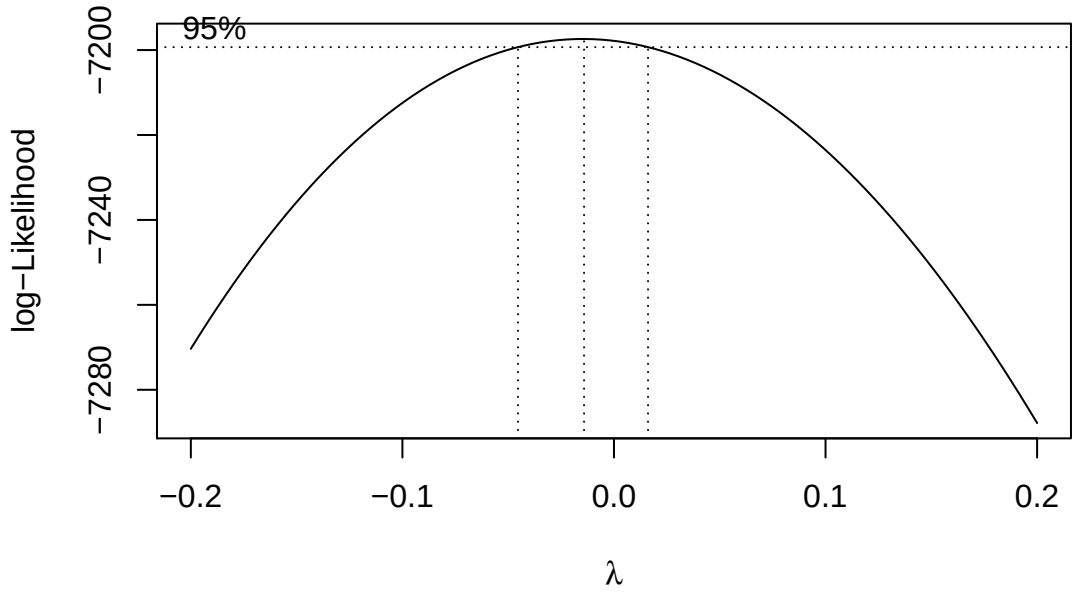


Figure 7: (#fig:f6.2)Box-Cox Plot

1.3.3 Square Root Transformed Predictors and Log Transformed Target

A plot of each numerical predictor square root transformed plotted against the log transformed `TARGET_AMT` as recommended by the Box-Cox plot still shows little improvement.

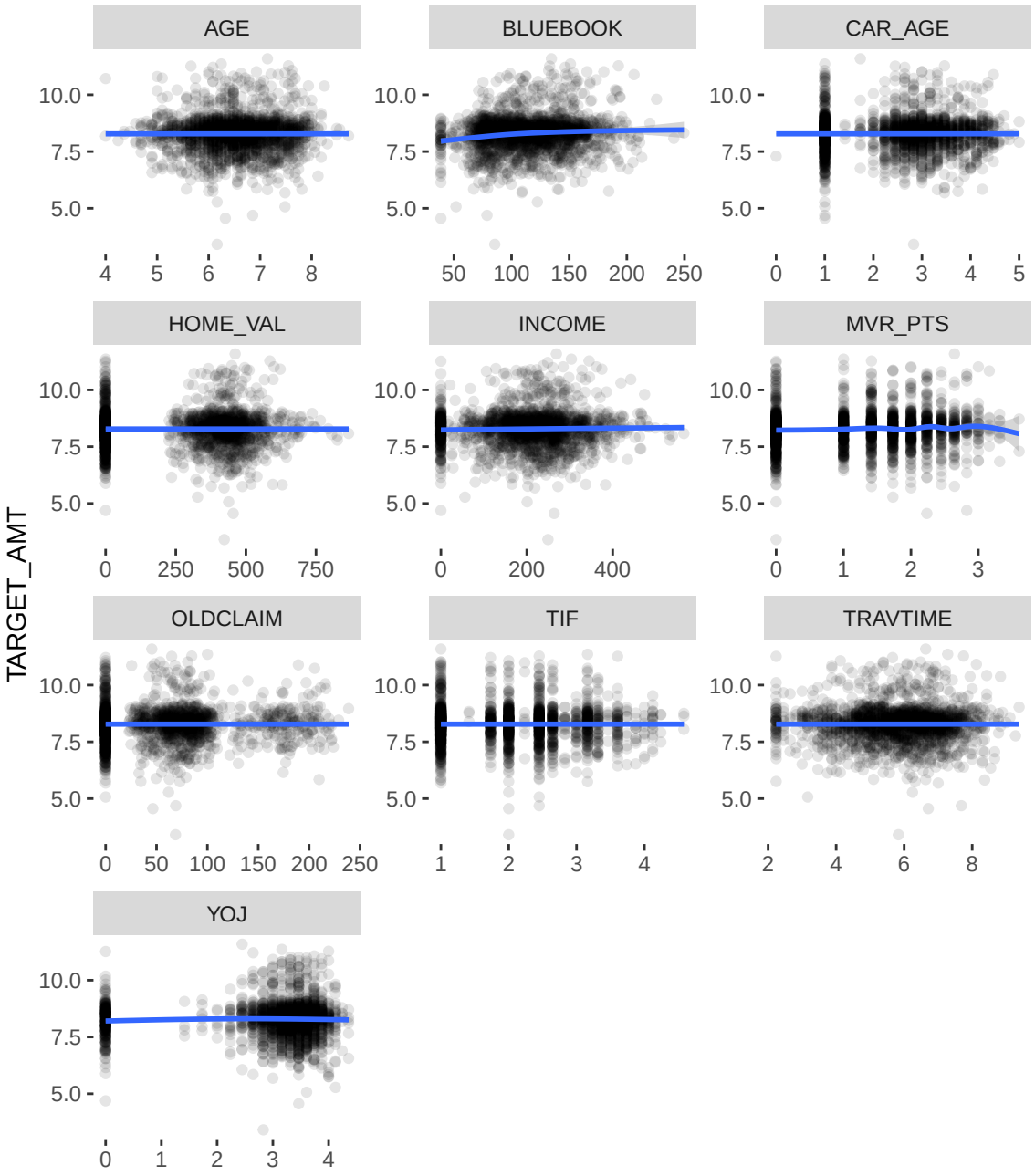


Figure 8: (#fig:f6.3)Scatter plot between square root transformed numeric predictors and the square root transformed TARGET_AMT filtered for rows where TARGET_AMT is greater than 0

1.4 Missing Data

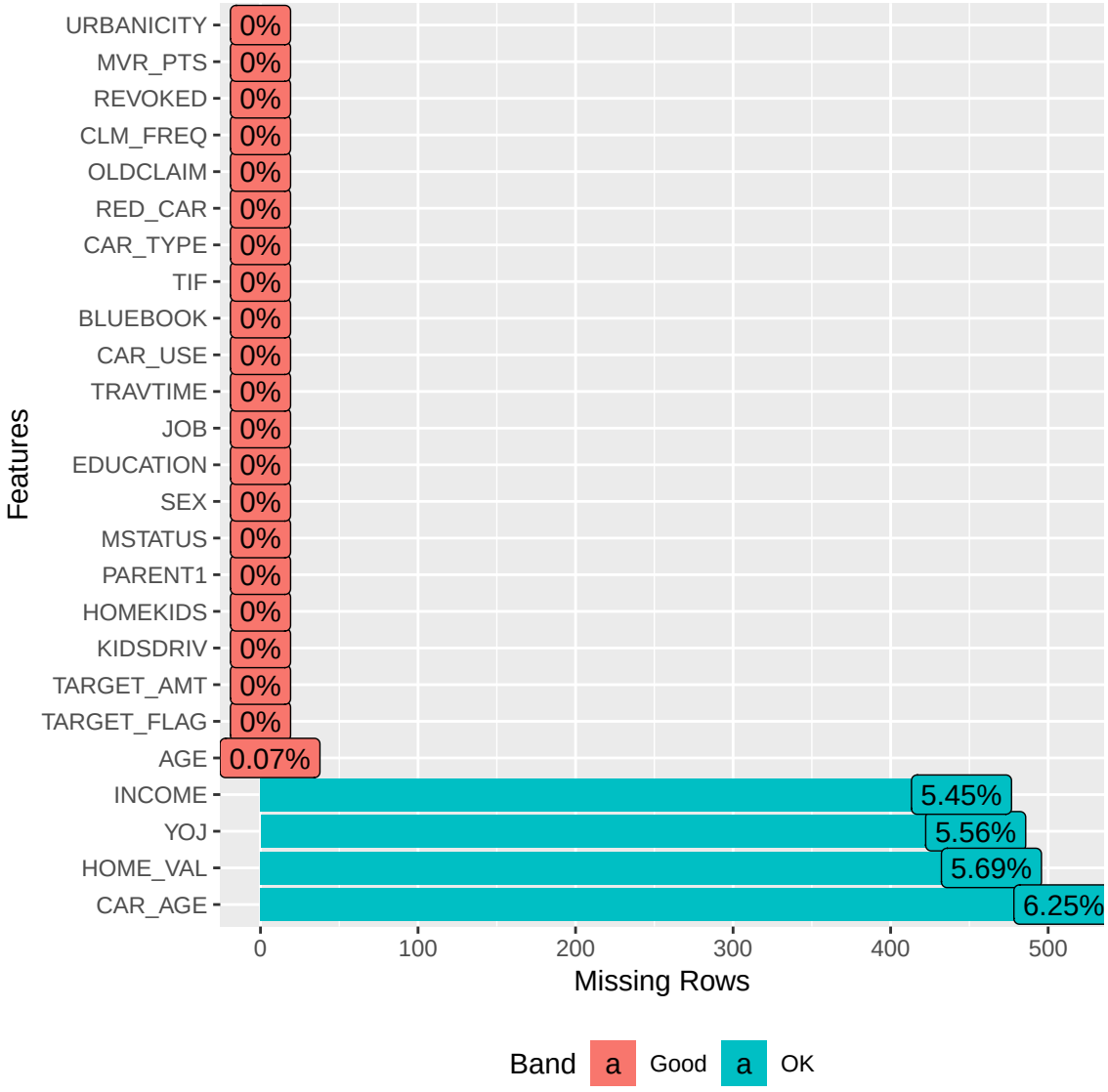


Figure 9: Missing data

A number of variables are missing observations: `AGE`, `INCOME`, `YOJ`, `HOME_VAL`, `CAR_AGE`. For `AGE`, the number is inconsequential, but the others range between 5% and 6% of total. Approximately 21% of the cases are missing one of these variables, and an additional 2% are missing more than one. For this reason, we don't suspect latent factors can account for the absences, and assume that these values are missing at random and can be imputed when preparing the data for modeling.

2 DATA PREPARATION

2.1 Missing Values

[JO: WHAT'S THE DIFFERENCE BETWEEN THE M AND MAXIT VALUES (1 FOR AGE, 2 FOR OTHERS?)] [GAB: CAN I KNOW THE ANSWER TO THIS TOO?]

To deal with missing data values for the variables `INCOME`, `YOJ`, `HOME_VAL`, and `CAR_AGE` - and to a lesser extent `AGE` - the MICE (Multivariate Imputation By Chained Equations) package was leveraged. The package assumes missing values are missing at random and creates multiple imputations (replacement values) for multivariate missing data using a a method based on Fully Conditional Specification, where each incomplete variable is imputed by a separate model. The method can impute mixes of continuous, binary, unordered categorical and ordered categorical, and continuous two-level data; and it can maintain consistency between imputations by means of passive imputation. The quality of imputed values was inspected using multiple diagnostic plots.

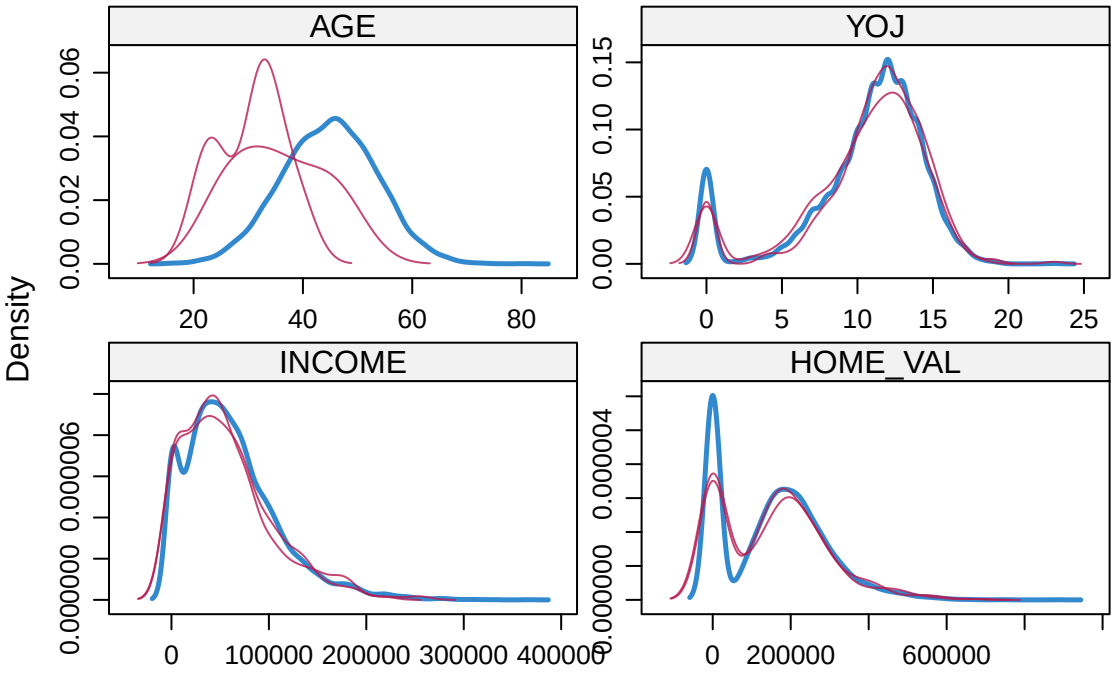
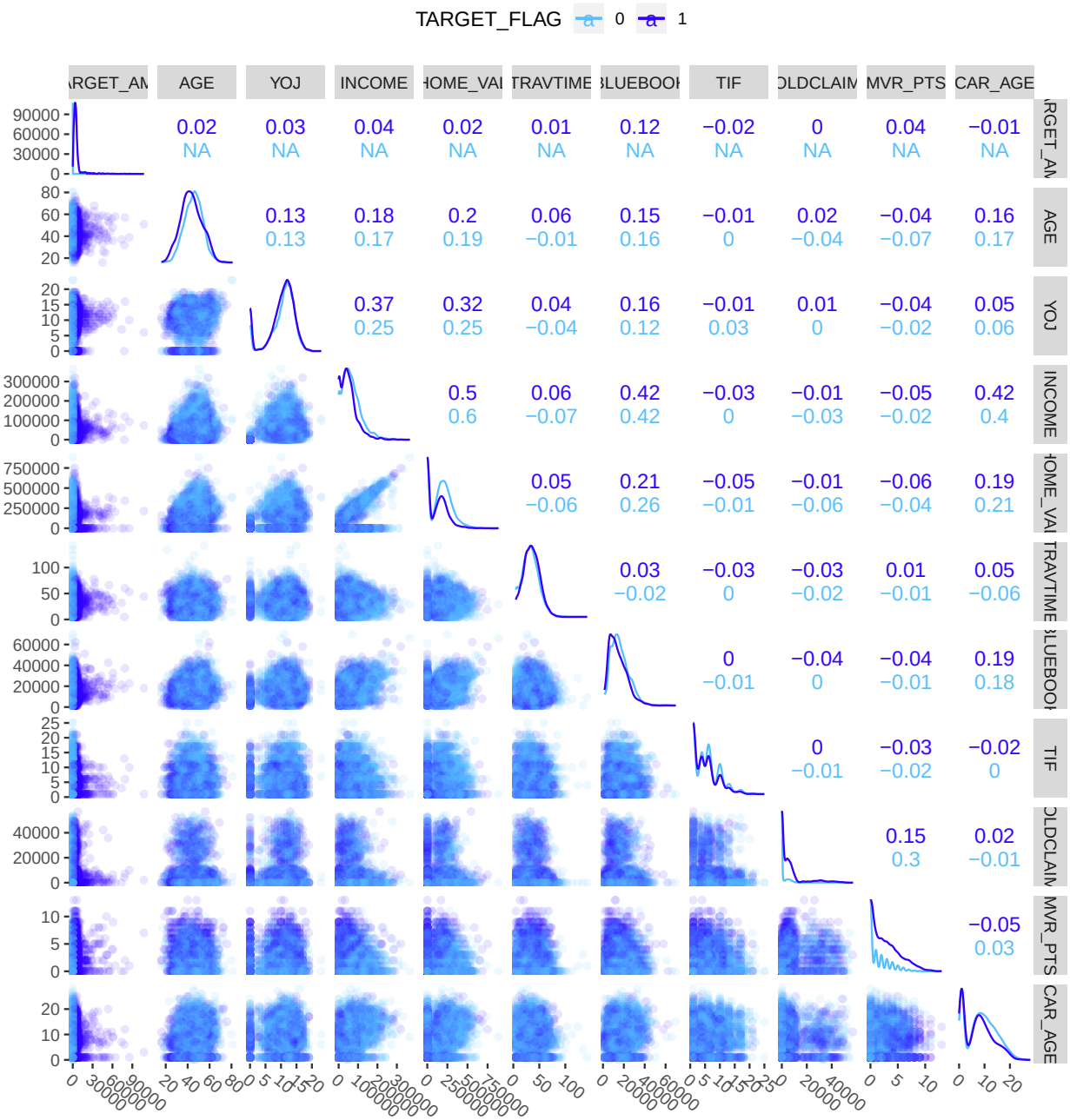


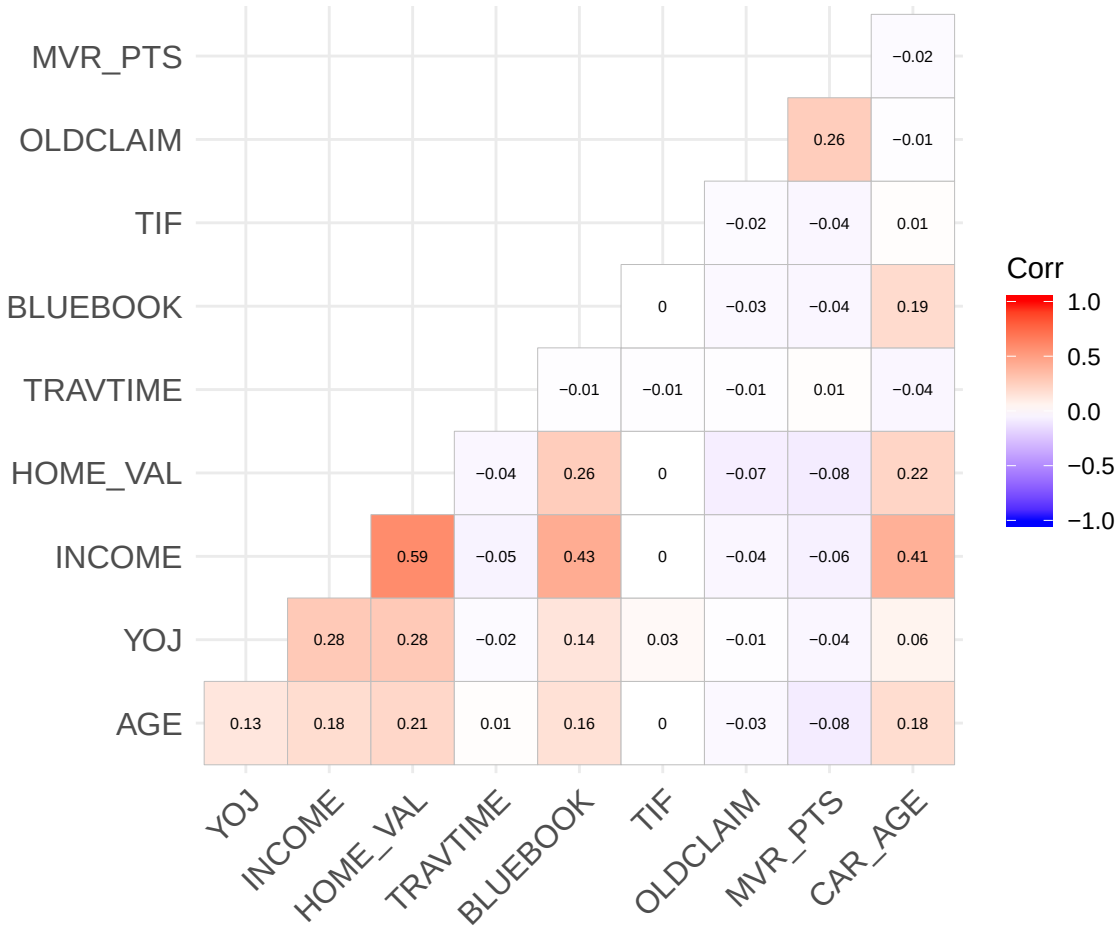
Figure 10: Difference between original and imputed data

The blue and red lines represent the distribution originally known and imputed values respectively. With the exception of `AGE`, the distribution of the imputed values accords with the distribution of pre-existing values. The imputed values will be used for the four variables. When it comes to `AGE` (only .07 or XX number of cases), it was to be imputed separately using median imputation.



Unsurprisingly, higher levels of `INCOME` are found with higher values of `YOJ`; this also means more income is disposable, which shows correlation with `HOME_VAL` and `BLUEBOOK`.

Additionally, `MVR_PTS` shows a positive correlation with `OLDCLAIMS`.


MVR_PTS
OLDCLAIM
TIF
BLUEBOOK
TRAVTIME
HOME_VAL
INCOME
YOJ
AGE
YOJ
INCOME
HOME_VAL
TRAVTIME
BLUEBOOK
TIF
OLDCLAIM
MVR_PTS
CAR_AGE
0.13
0.28 0.18
0.28 0.59 0.21
−0.02 −0.05 0.01
0.14 0.43 0.16
0.03 0 0 −0.01
−0.01 −0.04 −0.03 −0.01
−0.04 −0.06 −0.08 0.01
0.06 0.41 0.18 −0.04
−0.02 −0.01 0.01 0.19 −0.04 0.22
Corr
1.0
0.5
0.0
−0.5
−1.0

3 BUILD MODELS

3.1 Classification Models: Models 1, 2, 3

The first three models use the predictor variables in binary logistic models as inputs and interpret their contributions to predicting the likelihood of a claim. We use `drop` and `MASS:stepAIC` functions to judge which variables to remove, evaluating AIC statistics as we go.

3.1.1 Model 1 - Base model using categorical predictors only

TARGET_FLAG ~ PARENT1 + SEX + MSTATUS + EDUCATION + JOB + CAR_TYPE + CAR_USE + REVOKED + URBANICITY + KIDSDRIV + HOMEKIDS + CLM_FREQ

For an easily interpretable model aimed at predicting TARGET_FLAG, inputs for Model 1 were restricted to categorical variables alone. The AIC metric as well as the p-value and significance code suggests that the `RED_CAR` variable could be removed, so this predictor was removed from model 1. Model 1 serves as a base model from which to compare other models.

Observations	7651
Dependent variable	TARGET_FLAG
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(37)$	1764.49
Pseudo-R^2 (Cragg-Uhler)	0.30
Pseudo-R^2 (McFadden)	0.20
AIC	7125.59
BIC	7389.41

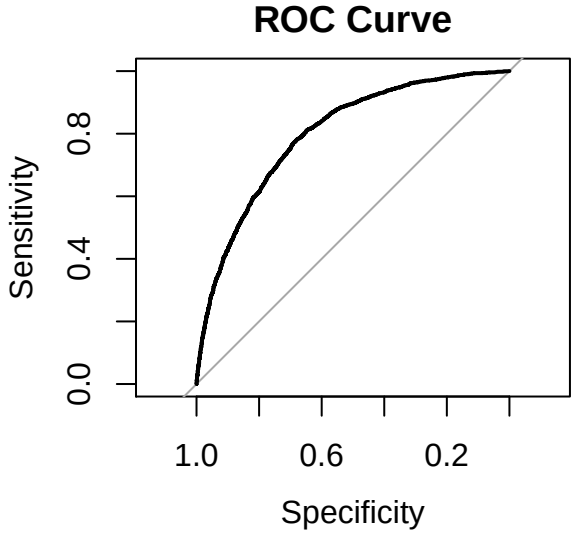


Figure 11: Model 1 ROC Curve

	Est.	S.E.	z val.	p	VIF
(Intercept)	-1.68	0.23	-7.27	0.00	NA
PARENT1Yes	0.22	0.12	1.78	0.07	2.34
SEXF	-0.31	0.09	-3.38	0.00	2.36
MSTATUSNo	0.68	0.08	8.96	0.00	1.67
EDUCATIONBachelors	-0.51	0.11	-4.72	0.00	7.49
EDUCATIONMasters	-0.41	0.16	-2.51	0.01	7.49
EDUCATIONPhD	-0.49	0.20	-2.51	0.01	7.49
EDUCATIONHigh School	-0.03	0.10	-0.35	0.73	7.49
JOBClerical	0.61	0.20	3.11	0.00	14.06
JOBDoctor	-0.27	0.26	-1.03	0.30	14.06
JOBHome Maker	0.72	0.20	3.63	0.00	14.06
JOBLawyer	0.16	0.17	0.91	0.36	14.06
JOBManager	-0.61	0.17	-3.50	0.00	14.06
JOBProfessional	0.24	0.18	1.35	0.18	14.06
JOBStudent	0.72	0.20	3.52	0.00	14.06
JOBBlue Collar	0.43	0.19	2.28	0.02	14.06
CAR_TYPEPanel Truck	0.21	0.14	1.49	0.14	3.71
CAR_TYPEPickup	0.59	0.10	5.79	0.00	3.71
CAR_TYPESports Car	1.22	0.12	9.84	0.00	3.71
CAR_TYPEVan	0.40	0.12	3.22	0.00	3.71
CAR_TYPESUV	0.95	0.10	9.06	0.00	3.71
CAR_USEPrivate	-0.74	0.09	-7.96	0.00	2.47
REVOKEDYes	0.74	0.08	9.10	0.00	1.01
URBANICITYRural	-2.20	0.11	-19.34	0.00	1.11
KIDSDRIV1	0.40	0.11	3.51	0.00	1.54
KIDSDRIV2	0.69	0.16	4.20	0.00	1.54
KIDSDRIV3	0.86	0.32	2.72	0.01	1.54
KIDSDRIV4	-11.10	204.58	-0.05	0.96	1.54
HOMEKIDS1	0.34	0.11	3.01	0.00	2.65
HOMEKIDS2	0.28	0.11	2.51	0.01	2.65
HOMEKIDS3	0.27	0.13	2.09	0.04	2.65
HOMEKIDS4	0.05	0.21	0.26	0.80	2.65
HOMEKIDS5	0.03	0.69	0.05	0.96	2.65
CLM_FREQ1	0.64	0.09	7.41	0.00	1.06
CLM_FREQ2	0.68	0.08	8.51	0.00	1.06
CLM_FREQ3	0.67	0.09	7.08	0.00	1.06
CLM_FREQ4	0.97	0.17	5.61	0.00	1.06
CLM_FREQ5	1.00	0.55	1.84	0.07	1.06

Standard errors: MLE

Table 5: Area Under the Curve

x
0.8

3.1.2 Marginal Model Plots

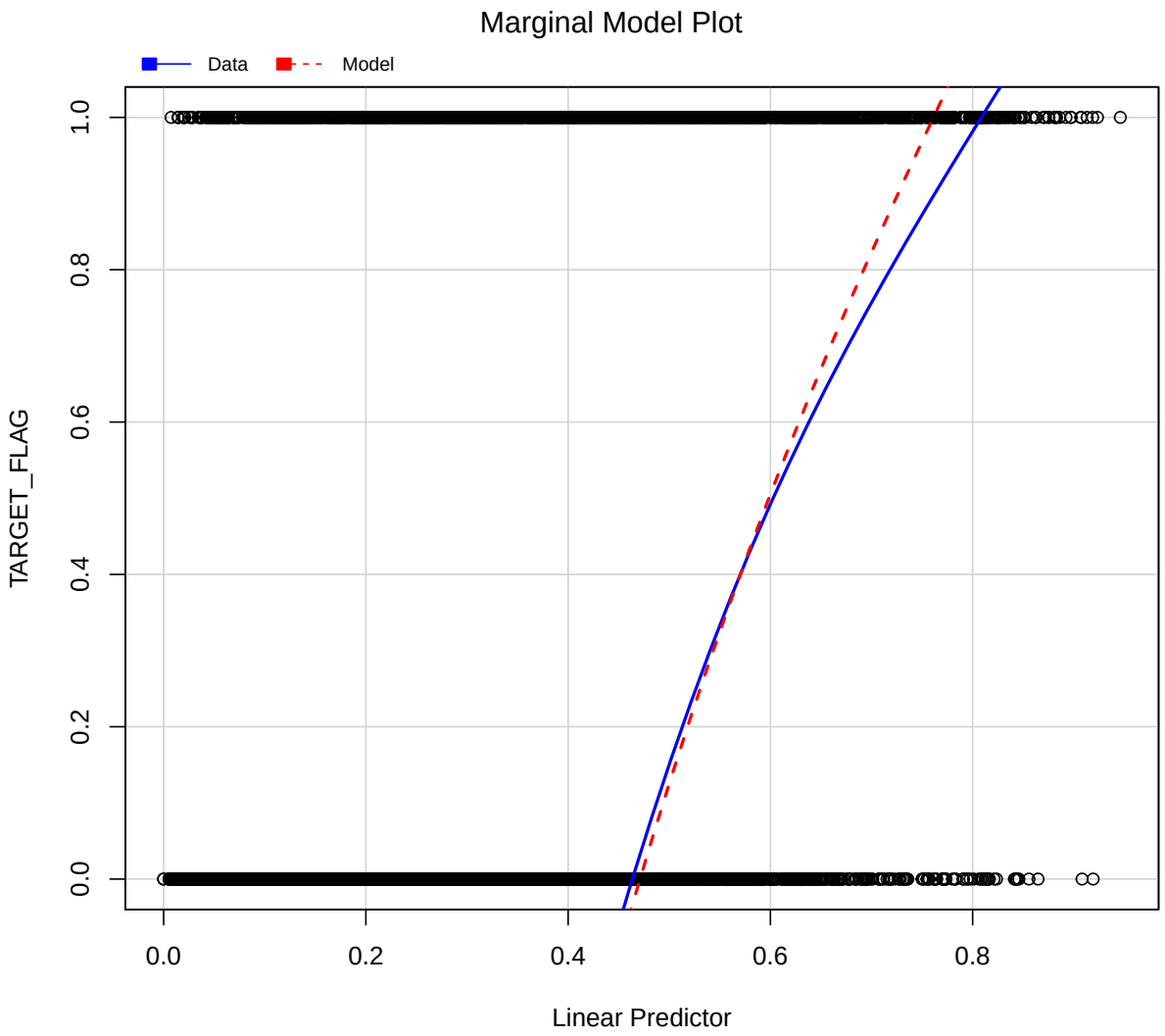


Figure 12: Model 1 Marginal Model Plots

3.1.3 Model 2 - Refined base model plus numerical predictors

TARGET_FLAG ~ MSTATUS + EDUCATION_Bachelors + JOB_Clerical + JOB_Manager + CAR_TYPE + CAR_USE + REVOKED + URBANICITY + KIDSDRIV + HOMEKIDS + CLM_FREQ + INCOME + HOME_VAL + TRAVTIME + BLUEBOOK + TIF + OLDCLAIM + MVR_PTS,

Building on model 1, model 2 excludes the `RED_CAR` variable and adds in all of the numerical variables to see if they add value to our model. After the initial model statistics were examined we then further refined the model by removing `AGE`, `CAR_AGE`, `SEX`, `YOJ`, and `PARENT1` due to low p-value significance. `EDUCATION`, and `JOB` were only significant if the education was 'Bachelors' or if the job was 'Manager', so two new binary variables, `EDUCATION_Bachelors` and `JOB_Manager` were added to the dataset indicating yes or no for these specific education and job values.

Observations	7651
Dependent variable	TARGET_FLAG
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(33)$	1986.84
Pseudo-R^2 (Cragg-Uhler)	0.33
Pseudo-R^2 (McFadden)	0.23
AIC	6895.24
BIC	7131.28

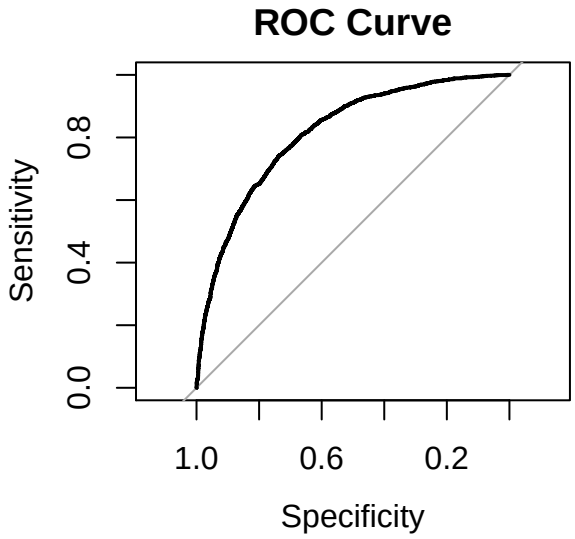


Figure 13: Model 3 ROC Curve

3.1.4 Marginal Model Plots

```
## Error in names(dat) <- object$term :
##   'names' attribute [1] must be the same length as the vector [0]
## Error in names(dat) <- object$term :
##   'names' attribute [1] must be the same length as the vector [0]
```

	Est.	S.E.	z val.	p	VIF
(Intercept)	-0.96	0.16	-6.18	0.00	NA
MSTATUSNo	0.64	0.08	8.55	0.00	1.55
EDUCATION_BachelorsTRUE	-0.27	0.07	-3.89	0.00	1.05
JOB_ClericalTRUE	0.26	0.09	3.05	0.00	1.16
JOB_ManagerTRUE	-0.77	0.11	-6.79	0.00	1.08
CAR_TYPEPanel Truck	0.52	0.15	3.51	0.00	2.23
CAR_TYPEPickup	0.46	0.10	4.50	0.00	2.23
CAR_TYPESports Car	0.93	0.11	8.47	0.00	2.23
CAR_TYPEVan	0.54	0.12	4.37	0.00	2.23
CAR_TYPESUV	0.67	0.09	7.61	0.00	2.23
CAR_USEPrivate	-0.90	0.07	-12.10	0.00	1.52
REVOKEDYes	0.95	0.10	9.96	0.00	1.36
URBANICITYRural	-2.29	0.12	-19.72	0.00	1.13
KIDSDRIV1	0.42	0.12	3.64	0.00	1.53
KIDSDRIV2	0.71	0.17	4.25	0.00	1.53
KIDSDRIV3	0.74	0.32	2.29	0.02	1.53
KIDSDRIV4	-11.57	199.91	-0.06	0.95	1.53
HOMEKIDS1	0.42	0.10	4.36	0.00	1.60
HOMEKIDS2	0.36	0.10	3.77	0.00	1.60
HOMEKIDS3	0.36	0.12	3.08	0.00	1.60
HOMEKIDS4	0.21	0.21	1.03	0.30	1.60
HOMEKIDS5	0.25	0.70	0.36	0.72	1.60
CLM_FREQ1	0.58	0.10	5.68	0.00	1.87
CLM_FREQ2	0.67	0.10	6.85	0.00	1.87
CLM_FREQ3	0.62	0.11	5.67	0.00	1.87
CLM_FREQ4	0.81	0.18	4.50	0.00	1.87
CLM_FREQ5	1.14	0.56	2.04	0.04	1.87
INCOME	-0.00	0.00	-5.97	0.00	1.84
HOME_VAL	-0.00	0.00	-3.34	0.00	1.94
TRAVTIME	0.01	0.00	7.67	0.00	1.04
BLUEBOOK	-0.00	0.00	-4.73	0.00	1.76
TIF	-0.06	0.01	-7.39	0.00	1.01
OLDCLAIM	-0.00	0.00	-4.62	0.00	1.89
MVR_PTS	0.10	0.01	6.83	0.00	1.24

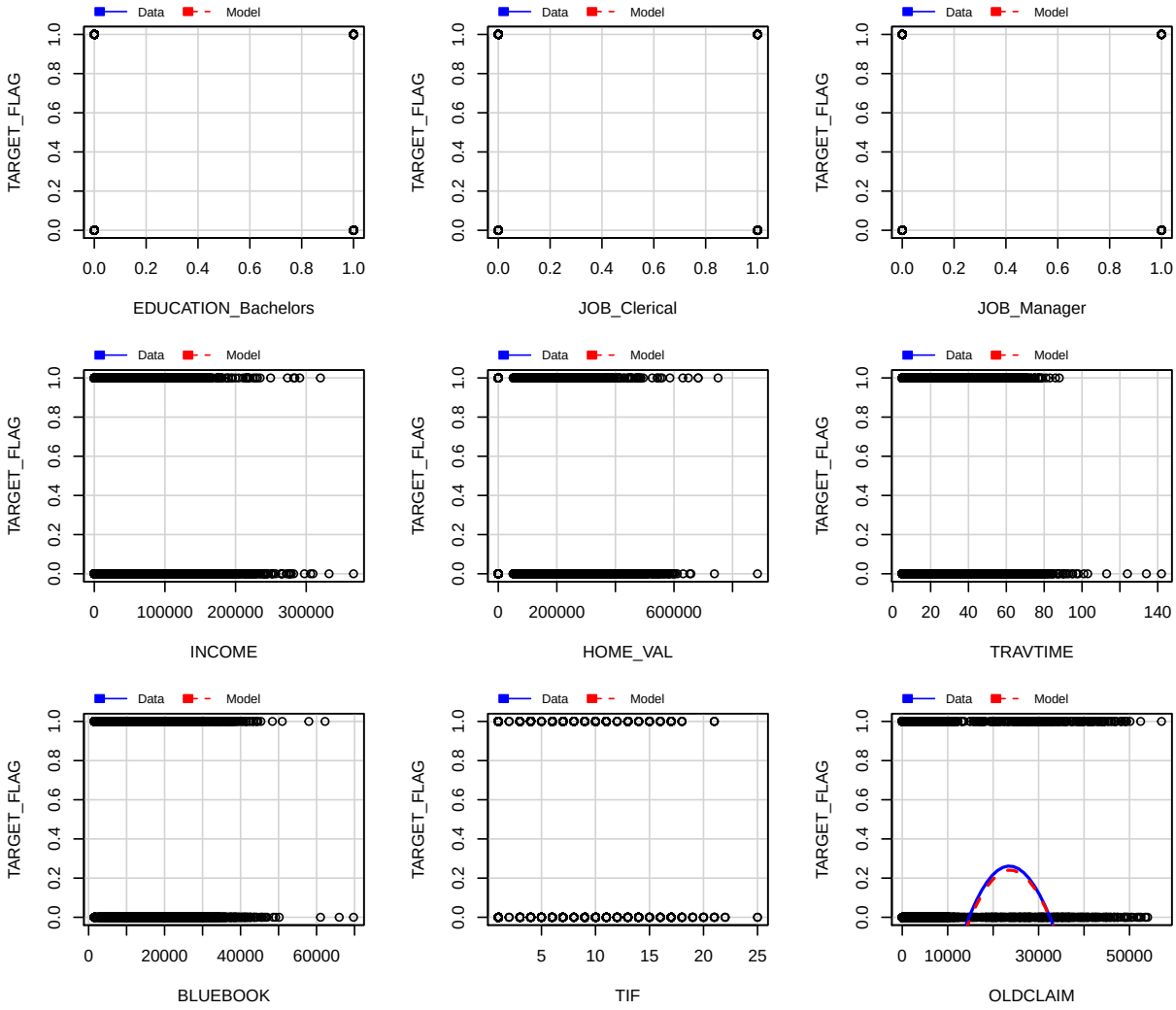
Standard errors: MLE

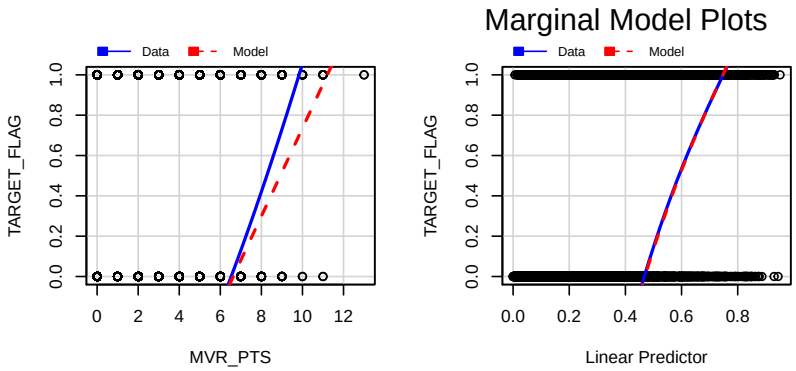
```
## Error in names(dat) <- object$term :
##   'names' attribute [1] must be the same length as the vector [0]
## Error in names(dat) <- object$term :
##   'names' attribute [1] must be the same length as the vector [0]
```

```
## Error in names(dat) <- object$term :
##   'names' attribute [1] must be the same length as the vector [0]
## Error in names(dat) <- object$term :
##   'names' attribute [1] must be the same length as the vector [0]
```

Table 6: Area Under the Curve

x
0.81




Marginal Model Plots
Data
Model
TARGET_FLAG
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
12
MVR_PTS
Data
Model
TARGET_FLAG
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
Linear Predictor

3.1.5 Model 3 - Binary logistic model

Model 3 takes a similar approach to model 2 by incorporating all numeric predictor variables plus the categorical predictors that were found to be singnificant in the previous model. Skewed numeric predictors (`BLUEBOOK`, `CAR_AGE`, `HOME_VAL`, `INCOME`, `MVR_PTS`, `OLDCLAIM`, `TIF`, and `TRAVTIME`,) were log transformed and added to the model as additional predictors. `AGE` and `YOJ` were not included since they were already normally distributed and determined not to be significant in the previous models. The model was then refined through backward elimination.

3.1.5.1 Original Model

TARGET_FLAG ~ MSTATUS + EDUCATION_Bachelors + JOB_Clerical + JOB_Manager + CAR_TYPE + CAR_USE + REVOKED + URBANICITY + KIDSDRIV + HOMEKIDS + CLM_FREQ + BLUEBOOK + CAR_AGE + HOME_VAL + INCOME + MVR_PTS + OLDCLAIM + TIF + TRAVTIME + log(BLUEBOOK) + log(CAR_AGE+1) + log(HOME_VAL+1) + log(INCOME+1) + log(MVR_PTS+1) + log(OLDCLAIM+1) + log(TIF) + log(TRAVTIME),

3.1.5.2 Model after backward elimination process

TARGET_FLAG ~ MSTATUS + EDUCATION_Bachelors + JOB_Clerical + JOB_Manager + CAR_TYPE + CAR_USE + REVOKED + URBANICITY + KIDSDRIV + HOMEKIDS + CAR_AGE + HOME_VAL + INCOME + MVR_PTS + OLDCLAIM + log(BLUEBOOK) + log(INCOME+1) + log(OLDCLAIM+1) + log(TIF) + log(TRAVTIME),

Observations	7651
Dependent variable	TARGET_FLAG
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(31)$	2017.05
Pseudo-R^2 (Cragg-Uhler)	0.34
Pseudo-R^2 (McFadden)	0.23
AIC	6861.02
BIC	7083.19

	Est.	S.E.	z val.	p	VIF
(Intercept)	1.13	0.57	1.96	0.05	NA
MSTATUSNo	0.66	0.08	8.79	0.00	1.55
EDUCATION_BachelorsTRUE	-0.26	0.07	-3.61	0.00	1.05
JOB_ClericalTRUE	0.28	0.09	3.01	0.00	1.30
JOB_ManagerTRUE	-0.73	0.11	-6.49	0.00	1.08
CAR_TYPEPanel Truck	0.45	0.14	3.17	0.00	1.96
CAR_TYPEPickup	0.46	0.10	4.56	0.00	1.96
CAR_TYPESports Car	0.90	0.11	8.12	0.00	1.96
CAR_TYPEVan	0.55	0.12	4.46	0.00	1.96
CAR_TYPESUV	0.67	0.09	7.64	0.00	1.96
CAR_USEPrivate	-0.89	0.08	-11.69	0.00	1.57
REVOKEDYes	0.97	0.10	10.10	0.00	1.37
URBANICITYRural	-2.31	0.12	-19.87	0.00	1.13
KIDSDRIV1	0.44	0.12	3.73	0.00	1.53
KIDSDRIV2	0.72	0.17	4.30	0.00	1.53
KIDSDRIV3	0.76	0.33	2.32	0.02	1.53
KIDSDRIV4	-11.62	190.55	-0.06	0.95	1.53
HOMEKIDS1	0.41	0.10	4.17	0.00	1.61
HOMEKIDS2	0.33	0.10	3.44	0.00	1.61
HOMEKIDS3	0.34	0.12	2.84	0.00	1.61
HOMEKIDS4	0.15	0.21	0.72	0.47	1.61
HOMEKIDS5	0.21	0.70	0.30	0.76	1.61
CAR_AGE	-0.02	0.01	-3.15	0.00	1.30
HOME_VAL	-0.00	0.00	-3.05	0.00	1.97
INCOME	-0.00	0.00	-3.08	0.00	2.57
MVR_PTS	0.10	0.01	6.74	0.00	1.24
OLDCLAIM	-0.00	0.00	-5.59	0.00	2.33
log(BLUEBOOK)	-0.32	0.06	-5.51	0.00	1.48
log(INCOME + 1)	-0.03	0.01	-2.82	0.00	1.75
log(OLDCLAIM + 1)	0.08	0.01	8.20	0.00	2.22
log(TIF)	-0.23	0.03	-7.26	0.00	1.01
log(TRAVTIME)	0.41	0.05	7.60	0.00	1.03

Standard errors: MLE

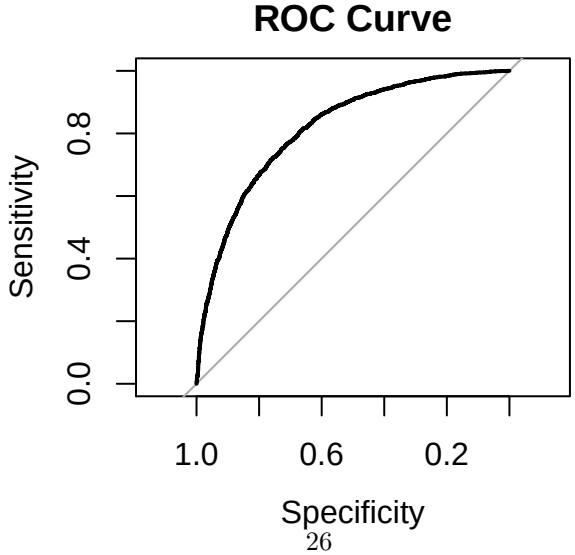


Figure 14: Model 4 ROC Curve

Table 7: Area Under the Curve

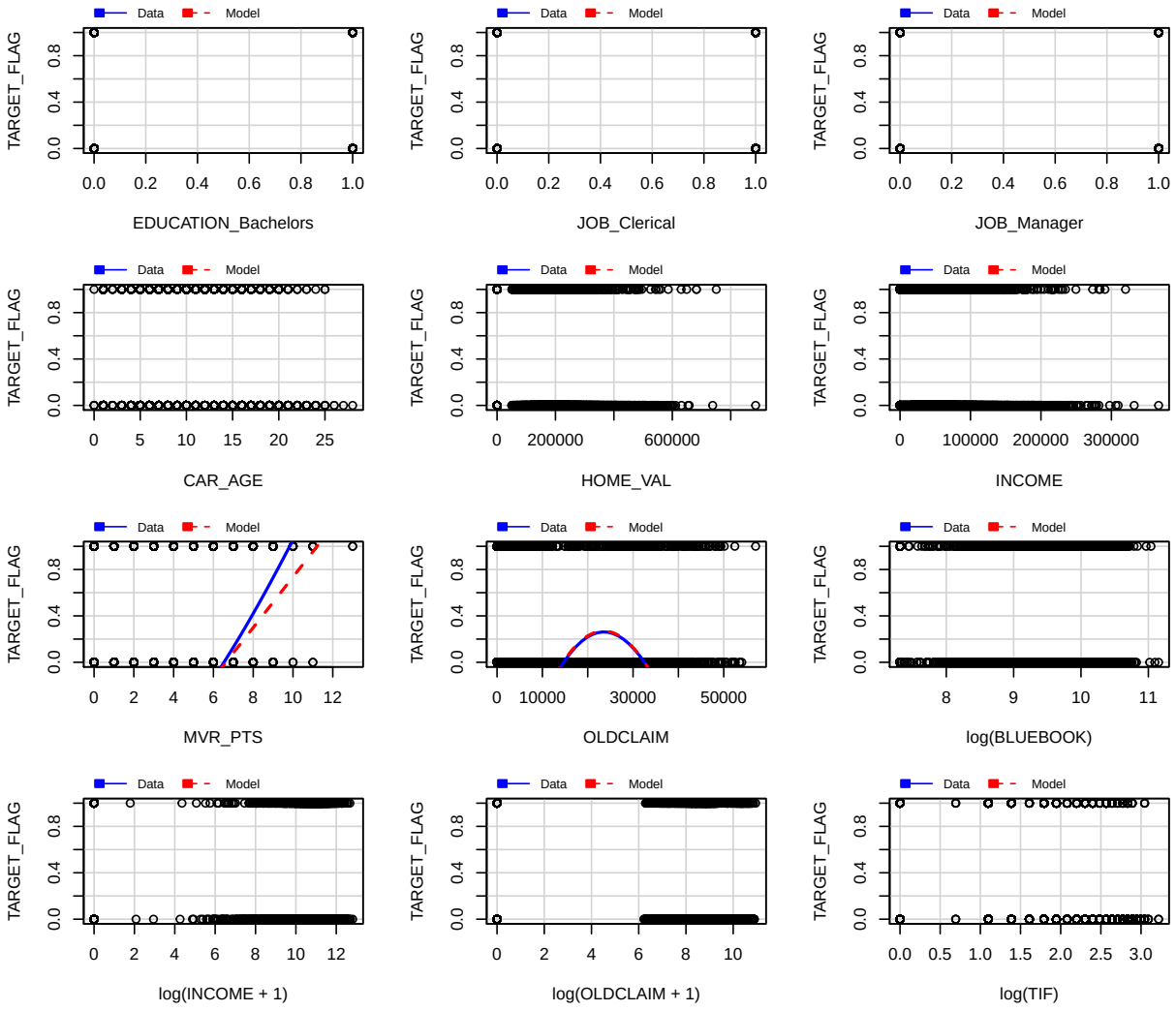
x
0.82

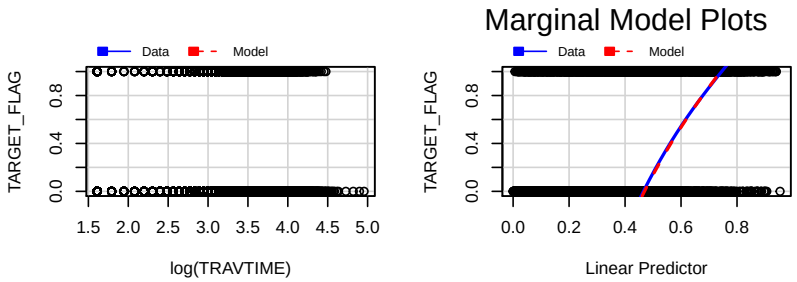
3.1.6 Marginal Model Plots

```
## Error in names(dat) <- object$term :
##   'names' attribute [1] must be the same length as the vector [0]
## Error in names(dat) <- object$term :
##   'names' attribute [1] must be the same length as the vector [0]
```

```
## Error in names(dat) <- object$term :
##   'names' attribute [1] must be the same length as the vector [0]
## Error in names(dat) <- object$term :
##   'names' attribute [1] must be the same length as the vector [0]
```

```
## Error in names(dat) <- object$term :
##   'names' attribute [1] must be the same length as the vector [0]
## Error in names(dat) <- object$term :
##   'names' attribute [1] must be the same length as the vector [0]
```


Data
Model
TARGET_FLAG
EDUCATION_Bachelors
JOB_Clerical
JOB_Manager
CAR_AGE
HOME_VAL
INCOME
MVR_PTS
OLDCLAIM
log(BLUEBOOK)
log(INCOME + 1)
log(OLDCLAIM + 1)
log(TIF)


Marginal Model Plots
Data
Model
TARGET_FLAG
0.0
0.4
0.8
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
log(TRAVTIME)
Data
Model
TARGET_FLAG
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
Linear Predictor

3.2 Regression Model: Models 5, 6

The next two models are multiple linear regression models aimed at predicting the value of claims based on different approaches, including constraining the cases based on `TARGET_FLAG` (i.e. based on whether or not a claim was filed) and different approaches to selecting explanatory variables.

23 lines were removed where `TARGET_AMT` greater than $45,000; these lines had a `BLUEBOOK` value far less than the car crash cost. A new variable `mileage` was created based on `TRAVTIME` and `CAR_AGE`.

3.2.1 Model 5 - Multiple linear regression model

Model 5 is a multiple linear regression model built only on cases with claims where `TARGET_FLAG` equals 1. The model is refined using stepwise elimination. From the model summary it can be observed that the Adjusted R-squared value is very low at 0.04.

Observations	1988
Dependent variable	TARGET_AMT
Type	OLS linear regression

F(49,1938)	1.46
R^2	0.04
Adj. R^2	0.01

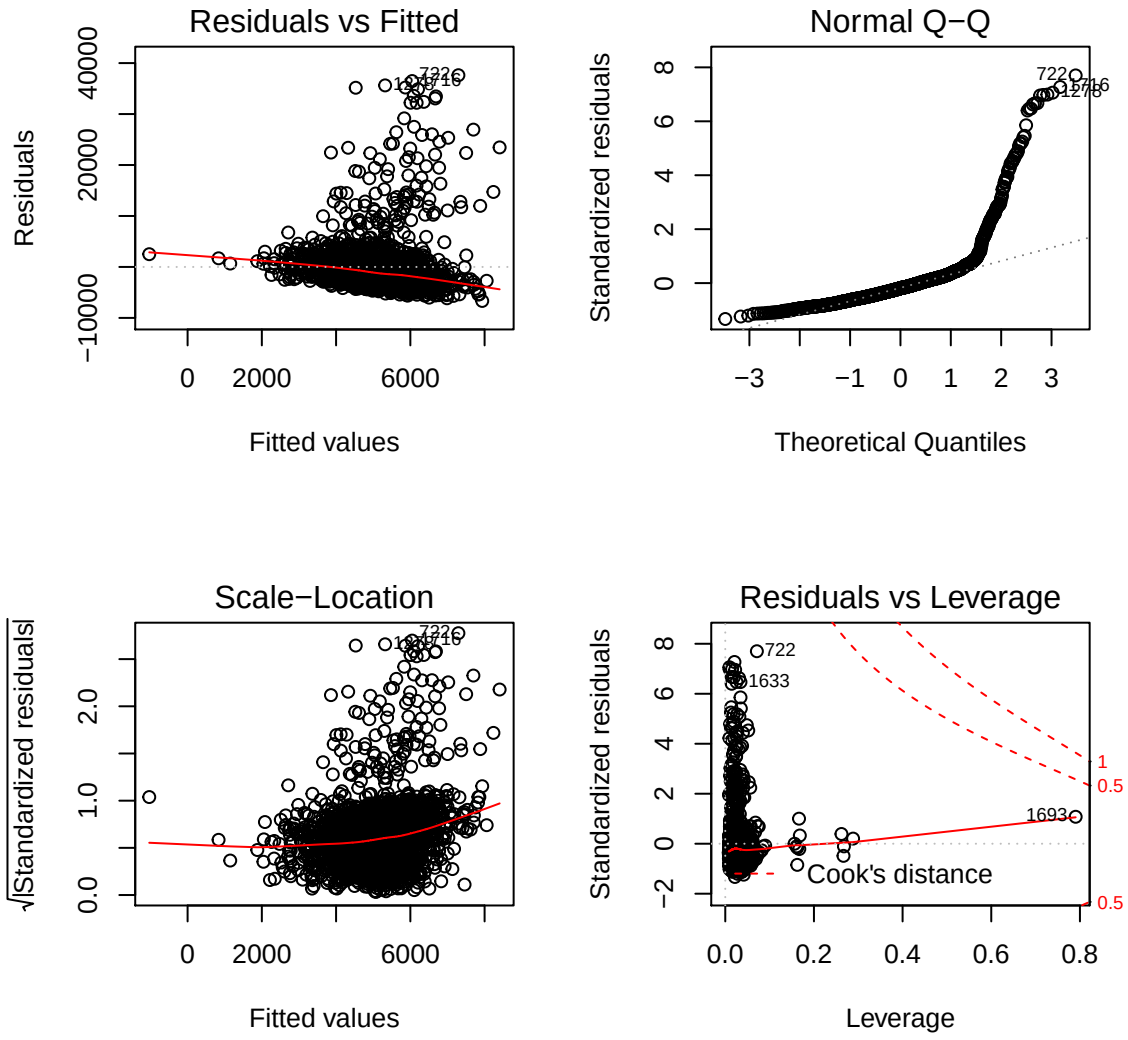


Figure 15: Model 5 Diagnostic Plots

3.2.2 Model 6 - Multiple linear regression model

Model 6 is a multiple linear regression model built on all cases - in other words, it relaxed the constraint that a claim was filed, and so includes `TARGET_AMT` values of 0. Forward elimination was used to refine variable selection. The Adjusted R-squared value significantly improved compared to the previous model.

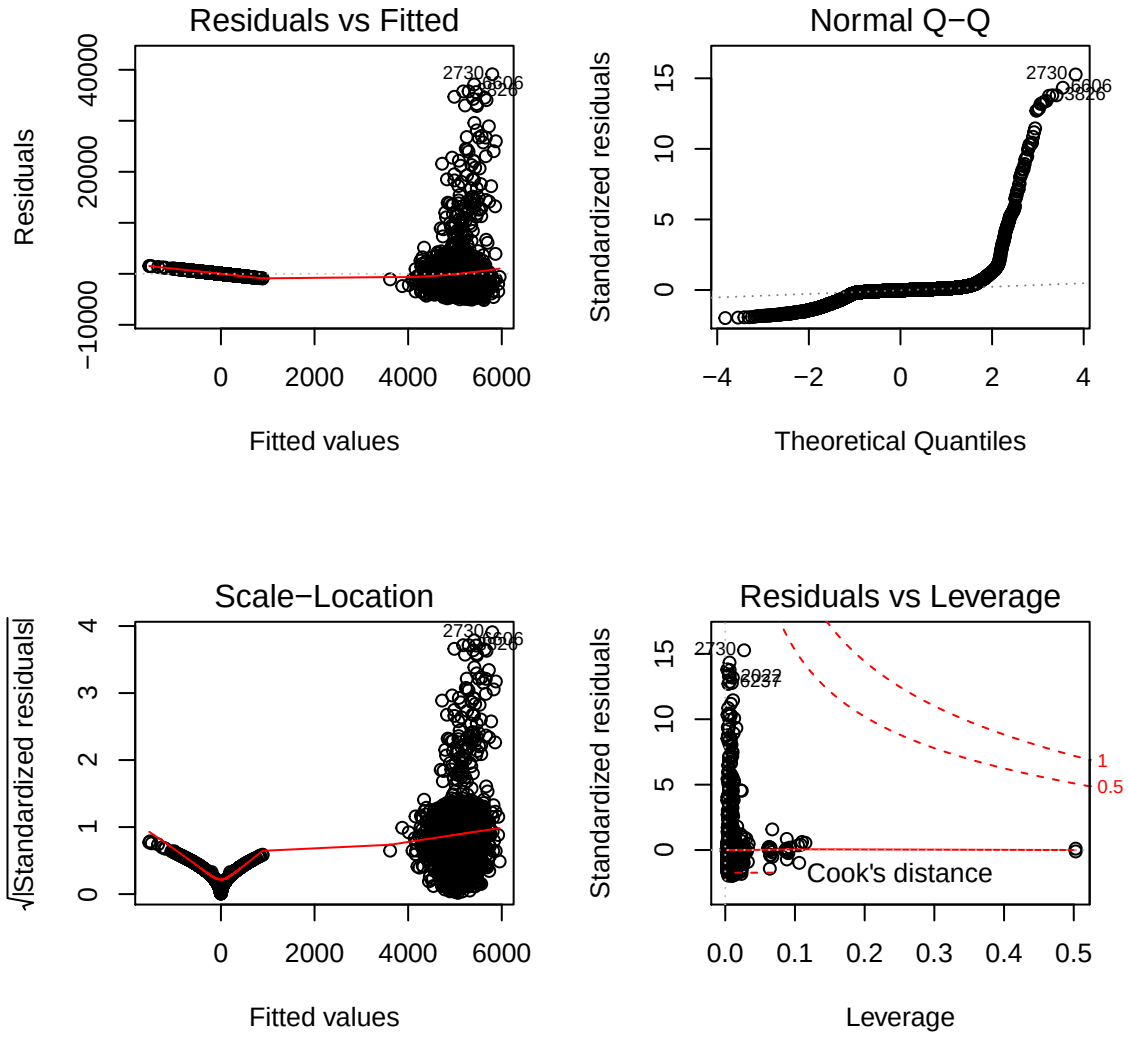


Figure 16: Model 6 Diagnostic Plots

4 SELECT MODELS

4.1 Pseudo R2

There is no R^2 for logistic regression to further evaluate, however, there is an alternative called $pseudoR^2$ terms that can be used for evaluation.

5 Appendix

The appendix is available as script.R file in `project4_insurance` folder.

https://github.com/betsyrosalen/DATA_621_Business_Analyt_and_Data_Mining

	Est.	S.E.	t val.	p	VIF
(Intercept)	13105.50	14676.12	0.89	0.37	NA
KIDSDRIV1	106.21	442.21	0.24	0.81	1.96
KIDSDRIV2	425.38	583.12	0.73	0.47	1.96
KIDSDRIV3	-257.74	1054.66	-0.24	0.81	1.96
log(AGE)	-4711.83	3445.76	-1.37	0.17	49.10
AGE	124.74	82.67	1.51	0.13	48.42
HOMEKIDS1	319.95	476.22	0.67	0.50	4.48
HOMEKIDS2	296.74	468.87	0.63	0.53	4.48
HOMEKIDS3	432.90	529.25	0.82	0.41	4.48
HOMEKIDS4	1207.11	821.24	1.47	0.14	4.48
HOMEKIDS5	942.71	2604.73	0.36	0.72	4.48
YOJ	-48.97	46.55	-1.05	0.29	3.36
log(INCOME + 0.00000000000001)	19.03	17.37	1.10	0.27	4.58
INCOME	-0.01	0.00	-1.67	0.10	2.93
CAR_AGE	-68.70	35.23	-1.95	0.05	2.94
log(mileage)	111.30	79.38	1.40	0.16	1.61
log(BLUEBOOK)	991.60	472.70	2.10	0.04	7.54
BLUEBOOK	-0.01	0.04	-0.20	0.84	10.21
TIF	-24.96	29.33	-0.85	0.39	1.03
log(OLDCLAIM + 0.00000000000001)	188.24	331.78	0.57	0.57	3476.34
OLDCLAIM	0.01	0.03	0.40	0.69	7.23
CLM_FREQ1	-7805.75	13351.39	-0.58	0.56	3694.41
CLM_FREQ2	-7745.34	13348.65	-0.58	0.56	3694.41
CLM_FREQ3	-8067.59	13354.06	-0.60	0.55	3694.41
CLM_FREQ4	-9164.28	13353.88	-0.69	0.49	3694.41
CLM_FREQ5	-8497.12	13546.44	-0.63	0.53	3694.41
MVR_PTS	69.29	48.96	1.42	0.16	1.23
PARENT1Yes	-192.34	463.54	-0.41	0.68	2.85
SEXF	-429.18	416.33	-1.03	0.30	3.31
EDUCATIONBachelors	295.14	445.00	0.66	0.51	12.56
EDUCATIONMasters	1180.38	755.96	1.56	0.12	12.56
EDUCATIONPhD	2571.09	918.26	2.80	0.01	12.56
EDUCATIONHigh School	-62.62	353.00	-0.18	0.86	12.56
JOBClerical	163.08	829.16	0.20	0.84	39.90
JOBDoctor	-1356.27	1187.50	-1.14	0.25	39.90
JOBHome Maker	-38.33	897.73	-0.04	0.97	39.90
JOBLawyer	424.71	701.95	0.61	0.55	39.90
JOBManager	-308.89	736.63	-0.42	0.68	39.90
JOBProfessional	126.37	774.22	0.16	0.87	39.90
JOBStudent	26.74	907.40	0.03	0.98	39.90
JOBBlue Collar	775.16	788.77	0.98	0.33	39.90
CAR_TYPEPanel Truck	277.72	684.21	0.41	0.68	8.57
CAR_TYPEPickup	419.01	413.46	1.01	0.31	8.57
CAR_TYPESports Car	599.78	519.48	1.15	0.25	8.57
CAR_TYPEVan	-47.23	532.48	-0.09	0.93	8.57
CAR_TYPESUV	638.25	462.22	1.38	0.17	8.57
REVOKEDYes	-603.17	361.54	-1.67	0.10	1.66
URBANICITYRural	-432.38	521.59	-0.83	0.41	1.05
MSTATUSNo	756.94	321.79	2.35	0.02	2.00
CAR_USEPrivate	-67.81	360.16	-0.19	0.85	2.51

Standard errors: OLS

Observations	7628
Dependent variable	TARGET_AMT
Type	OLS linear regression

F(51,7576)	112.78
R^2	0.43
Adj. R^2	0.43

	Est.	S.E.	t val.	p	VIF
(Intercept)	2191.67	4527.86	0.48	0.63	NA
TARGET_FLAG	5062.37	77.46	65.35	0.00	1.31
KIDSDRIV1	57.56	128.25	0.45	0.65	1.67
KIDSDRIV2	216.37	181.01	1.20	0.23	1.67
KIDSDRIV3	-78.46	359.75	-0.22	0.83	1.67
KIDSDRIV4	-616.44	1857.51	-0.33	0.74	1.67
log(AGE)	-1396.62	1071.81	-1.30	0.19	53.59
AGE	36.01	24.97	1.44	0.15	52.53
HOMEKIDS1	45.75	123.87	0.37	0.71	3.28
HOMEKIDS2	-11.80	121.21	-0.10	0.92	3.28
HOMEKIDS3	42.75	142.17	0.30	0.76	3.28
HOMEKIDS4	329.08	235.37	1.40	0.16	3.28
HOMEKIDS5	353.06	758.81	0.47	0.64	3.28
YOJ	-10.11	11.22	-0.90	0.37	2.38
log(INCOME + 0.00000000000001)	6.14	4.47	1.37	0.17	3.13
INCOME	-0.00	0.00	-1.66	0.10	2.59
CAR_AGE	-20.00	9.91	-2.02	0.04	3.61
log(mileage)	48.09	32.61	1.47	0.14	2.23
log(BLUEBOOK)	359.64	126.73	2.84	0.00	7.16
BLUEBOOK	-0.01	0.01	-0.93	0.35	8.94
TIF	-4.70	7.24	-0.65	0.52	1.02
log(OLDCLAIM + 0.00000000000001)	68.11	102.72	0.66	0.51	4753.49
OLDCLAIM	-0.00	0.01	-0.04	0.97	7.21
CLM_FREQ1	-2816.92	4132.78	-0.68	0.50	4698.69
CLM_FREQ2	-2790.20	4129.68	-0.68	0.50	4698.69
CLM_FREQ3	-2913.99	4131.39	-0.71	0.48	4698.69
CLM_FREQ4	-3347.18	4139.71	-0.81	0.42	4698.69
CLM_FREQ5	-3187.07	4191.07	-0.76	0.45	4698.69
MVR_PTS	33.53	15.99	2.10	0.04	1.33
PARENT1Yes	34.95	128.82	0.27	0.79	2.14
SEXF	-93.04	96.57	-0.96	0.34	2.62
EDUCATIONBachelors	66.15	121.79	0.54	0.59	11.39
EDUCATIONMasters	235.53	180.20	1.31	0.19	11.39
EDUCATIONPhD	483.03	214.94	2.25	0.02	11.39
EDUCATIONHigh School	-11.92	101.47	-0.12	0.91	11.39
JOBClerical	-50.35	203.12	-0.25	0.80	30.21
JOBDoctor	-253.93	242.30	-1.05	0.29	30.21
JOBHome Maker	-26.14	223.12	-0.12	0.91	30.21
JOBLawyer	85.76	175.67	0.49	0.63	30.21
JOBManager	-61.15	171.65	-0.36	0.72	30.21
JOBProfessional	-38.28	183.24	-0.21	0.83	30.21
JOBStudent	-94.47	227.54	-0.42	0.68	30.21
JOBBlue Collar	142.24	191.48	0.74	0.46	30.21
CAR_TYPEPanel Truck	130.49	168.40	0.77	0.44	6.06
CAR_TYPEPickup	103.81	101.14	1.03	0.30	6.06
CAR_TYPESports Car	149.94	130.06	1.15	0.25	6.06
CAR_TYPEVan	13.69	126.08	0.11	0.91	6.06
CAR_TYPESUV	132.04	106.64	1.24	0.22	6.06
REVOKEDYes	-132.70	104.51	-1.27	0.20	1.33
URBANICITYRural	-22.84	85.38	-0.27	0.79	1.34
MSTATUSNo	161.54	76.98	2.10	0.04	1.61
CAR_USEPrivate	0.53	97.43	0.01	1.00	2.51

Standard errors: OLS

Table 8: Confusion Matrix Summary Statistics

	Sensitivity	Specificity	Precision	Recall	F1
Model.1	0.97	0.20	0.77	0.97	0.86
Model.2	0.97	0.25	0.78	0.97	0.87
Model.3	0.97	0.26	0.79	0.97	0.87

Table 9: Pseudo R2

	llh	llhNull	G2	McFadden	r2ML	r2CU
Model.1	-3525	-4407	1764	0.20	0.21	0.30
Model.2	-3414	-4407	1987	0.23	0.23	0.33
Model.3	-3399	-4407	2017	0.23	0.23	0.34